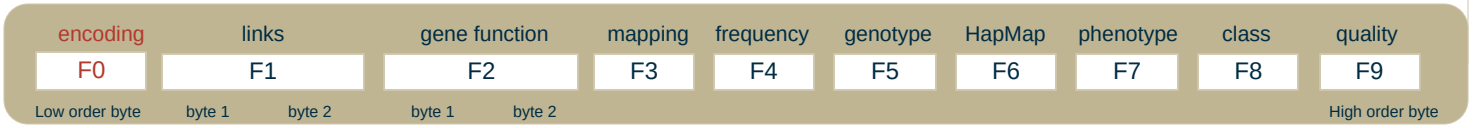
NCBI dbSNP bitfield encoding schema

Revisions:
Ver 5.4 Jun 2, 2010; Ver 5.3, Jul 16, 2009; Ver 5.2 - May 12, 2009; ver 5.1, Rev. Nov.5, 2008; ver 5, Rev. May 21 2008
ver 4, Rev. May 5 2008

12 byte structure



The dbSNP bitfield structure is a 12-byte object that defines variation properties in 9 areas.

- F1 and F2 are 2-byte structures
- F0 (lowest order byte) is the version of the encoding schema used for the data, see page 2 for definition.
- Bits labeled in **bold** are currently populated by dbSNP. Red color indicates newly implemented.

F1 – resource link properties

byte 2

8	*Reserve*
7	**SNP is Clinical(LSDB,OMIM,TPA,Diagnostic)**
6	**SNP is Precious(Clinical,Pubmed Cited)**
5	**Provisional Third Party Annotation(TPA)(currently rs from PHARMGKB who will give pheotype data).**
4	**Links exist to PubMed Central article**
3	Links exist to Short Read Archive
2	Has OrganismDBLink *(Ex. Jackson Lab for mouse)*
1	From MGC clone *(~20K rs set from specific submitter handle/ batch_id)*

byte 1

8	Links exist to Trace Archive
7	Links exist to Assembly Archive
6	Links exist to Entrez GEO
5	Links exist to ProbeDB
4	Links exist to Entrez Gene
3	Links exist to Entrez STS
2	**Has 3D structure** *SNP3D table*
1	**Has SubmitterLinkOut** *From SNP->SubSNP->Batch.link_out*

F2 – gene function properties

byte 2

8	*Reserve*
7	
6	**Has STOP-Loss** *A coding region variation where one allele in the set changes the encoded STOP CODON (TER). FxnClass = 43*
5	**Has non-synonymous frameshift** *A coding region variation where one allele in the set changes all downstream amino acids. FxnClass = 44*
4	**Has non-synonymous missense** *A coding region variation where one allele in the set changes protein peptide. FxnClass = 42*
3	**Has STOP-Gain** *A coding region variation where one allele in the set changes to STOP codon (TER). FxnClass = 41*
2	**Has reference** *A coding region variation where one allele in the set is identical to the reference sequence. FxnCode = 8*
1	**Has synonymous** *A coding region variation where one allele in the set does not change the encoded amino acid. FxnCode = 3*

byte 1

8	**In 3' UTR** *Location is in an untranslated region (UTR). FxnCode = 53*
7	**In 5' UTR** *Location is in an untranslated region (UTR). FxnCode = 55*
6	**In acceptor splice site** *FxnCode = 73*
5	**In donor splice-site** *FxnCode = 75*
4	**In Intron** *FxnCode = 6*
3	**In 3' gene region** *FxnCode = 13*
2	**In 5' gene region** *FxnCode = 15*
1	In gene segment *Defined as sequence intervals covered by a gene ID but not having an aligned transcript. FxnCode = 11*

F3 – mapping properties

8	*Reserve*
7	
6	
5	**Has other snp** *with exactly the same set of mapped positions on NCBI refernce assembly.*
4	**Has Assembly conflict.** *This is for weight 1 and 2 snp that maps to different chromosomes on different assemblies.*
3	**Is Assembly specific.** *This bit is 1 if the snp only maps to one assembly.*
2	**Weight (2 bits). Weight on NCBI reference assembly.** *map weight (2-bit: 00 = unmapped, 01 = 1, 10 = 2, 11 = 3 or more).*
1	

F4 – allele frequency properties

8	*Reserve*
7	
6	
5	
4	**Is mutation** (journal citation, explicit fact): *a low frequency variation that is cited in journal and other reputable sources.*
3	**Is Validated.** *This bit is set if the snp has 2+ minor allele count based on frequency or genotype data.*
2	**>5% minor allele frequency in each and all populations.**
1	**>5% minor allele frequency in 1+ populations**

NCBI dbSNP bitfield encoding schema

Ver 5.3, Jul 16, 2009; Ver 5.4, Jan 19,2011

F5 – genotype properties

8 7 6 5 4 — *Reserve*

3 — **Marker is on high density genotyping kit** *(50K density or greater). The snp may have phenotype associations present in dbGaP.*

2 — In Haplotype tagging set

1 — **Genotypes available.** *The snp has individual genotype (in SubInd table).*

F6 – Validation by HapMap/ TGP properties

8 — *Reserve*

7 — TGP 2010 production (for data created and released prior to ASHG)

6 — TGP_validated (for subset that passed positive second platform validation)

5 — TGP 2010 pilot (phases 1, 2, 3)

4 — TGP 2009 pilot phase 1

3 — Phase 3 genotyped: filtered, non-redundant.

2 — **Phase 2 genotyped:** filtered, non-redundant.

1 — Phase 1 genotyped: filtered, non-redundant.

F7 – phenotype properties

8 — Has MeSH is linked to a disease.

7 — **Variation is interrogated in a clinical diagnostic assay**

6 — Has transcription factor

5 — **Submitted from a locus-specific database.**

4 — Has p-value ≤ 1×10^{-3} in a dbGaP study association test

3 — Has LOD score ≥ 2.0 in a dbGaP study genome scan

2 — **Microattribution/third-party annotation(TPA:GWAS,PAGE)**

1 — **Has OMIM/OMIA**

F8 – variation class

8 7 6 5 — *Reserve*

4 3 2 1 —
0001 = single base polymorphism SNP
0010 = dips (deletion/insertion)
0011 = HETEROZYGOUS
0100 = Microsatellite
0101 = Named variation, e.g. (Alu)
0110 = NOVARIATION
0111 = mixed class
1000 = multi-base polymorphism

F9 – quality check

8 — *Reserve*

7 — Is suspect. The variants are *paralogous sequence differences. (added 01/19/11 ver 5.4) val=64*

6 — Variation is somatic, not germline. *The variation was detected in a Somatic tissue (e.g. cancer tumor). The variation is not known to exist in heritable DNA.*

5 — **Contig allele not present in SNP allele list**. *The reference sequence allele at the mapped position is not present in the SNP allele list, adjusted for orientation.*

4 — **Is Withdrawn by submitter** *If one member ss is withdrawn by submitter, then this bit is set. If all member ss' are withdrawn, then the rs is deleted to SNPHistory.*

3 — **Rs cluster has non-overlapping allele sets**. *True when rs set has more than 2 alleles from different submissions and these sets share no alleles in common.*

2 — Is a strain-specific fixed difference

1 — **Has Genotype Conflict** *Same (rs, ind), different genotype. N/N is not included.*

F0 – Version encoding

8 7 6 5 4 — *Reserve*

3 2 1 — **Bitmap schema version.** *Versions increment as integer value (current is version 2, version 1.2 is encoded as version 1)*